BBN TECHNOLOGIES
A Verizon Company

EARS The Blue Team
BBN LIMSI UPitt UW

Rich Transcription Evaluation Framework

Francis Kubala, Amit Srivastava, Daben Liu

EARS RT03-F Workshop
13 November 2003

Overview

- The Rich Transcription (RT) Framework
  - Token Error Rate for overall RT evaluation
  - Slot Error Rate for individual Metadata evaluation

- *rteval* - a tool for RT evaluation
  - Description
  - Demonstration

Brief History

- RT Framework first proposed at January workshop
  - White paper distributed in February
- First *rteval* version demonstrated in April
- Data structures defined in subcommittee, June-July
  - Included both RT Framework (RT XML) and NIST tools (RTTM)
- RT03-F Tasks clarified in August
  - Included substantial rewriting of Evaluation Plan
- Final *rteval* version released in September
  - Supports all of the RT03-F MDE Tasks

Most of BBN's MDE effort in 2003 was devoted to defining and implementing the RT Framework

BBN TECHNOLOGIES
A Verizon Company

What is the RT Framework?

Straightforward extension of the STT evaluation framework

RT					
STT	Metadata				
lexeme identity	filler	edit	IP	SU end	speaker label
he		x	x		1
he's					1
really					1
uh	x		x		1
out					1
of					1
line				x	1
yeah					2
right				x	2

an RT Token is
a STT Lexeme
+ associated Metadata

9 RT Tokens

BBN TECHNOLOGIES
A Verizon Company

Token Error Rate (TER) for RT

- The reference RT is an ordered sequence of RT Tokens
  - RT systems attempt to reproduce the reference RT Token sequence exactly
  - The sequence of system output tokens is aligned to the reference token sequence

- TER is the primary performance metric in the RT Framework
  - TER considers all token attributes jointly
  - An error in any attribute (or any number of attributes) counts as 1 Token Error

$$\text{TER} = \frac{100 * (\#\text{sub} + \#\text{del} + \#\text{ins})}{\#\text{ reference tokens}}$$

reference tokens = # STT lexemes

BBN TECHNOLOGIES
A Verizon Company

Modified DP Alignment for RT

- As in STT evaluation, Dynamic Programming is used to align the token sequences
  - For STT, only the lexeme identity is used for alignment
  - For RT, lexeme identity + metadata attributes control alignment, but with constraints:
  3. Metadata cannot prevent matching lexemes from aligning
  4. Metadata determines the alignment whenever lexeme identities are mismatched

- DP constraints are implemented as a simple table of token substitution costs
  - The substitution cost for (any number of) metadata errors for a token is less than the cost of a mismatch in lexeme identity

STT WER is preserved in the RT TER alignment

Caveat

- For RT03-F, *rteval* computes a baseline lexeme error called, RT1 (should be called Lexeme Error Rate)
  - RT1 is not equivalent to STT
  - Word fragments and filled pauses are not optionally deletable in RT1

Slot Error Rate (SER) for Metadata

- The RT Framework defines a SER for each metadata type
  - Useful for demonstrating improvement in metadata subsystem performance

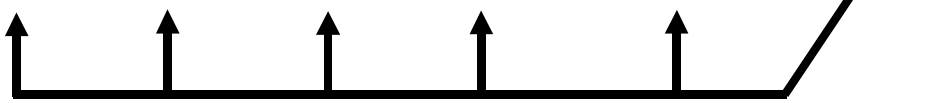
$$SER = \frac{100 * (\#sub + \#del + \#ins)}{\#\ reference\ slots}$$

reference slots differs for each metadata type

- Metadata SER is computed from the same alignment used for TER
  - SER is affected by underlying STT error, both in practice and in scoring

Metadata SER in Context

RT					
STT	Metadata				
lexeme identity	filler	edit	IP	SU end	speaker label
he		x	x		1
he's					1
really					1
uh	x		x		1
out					1
of					1
line				x	1
yeah					2
right				x	2



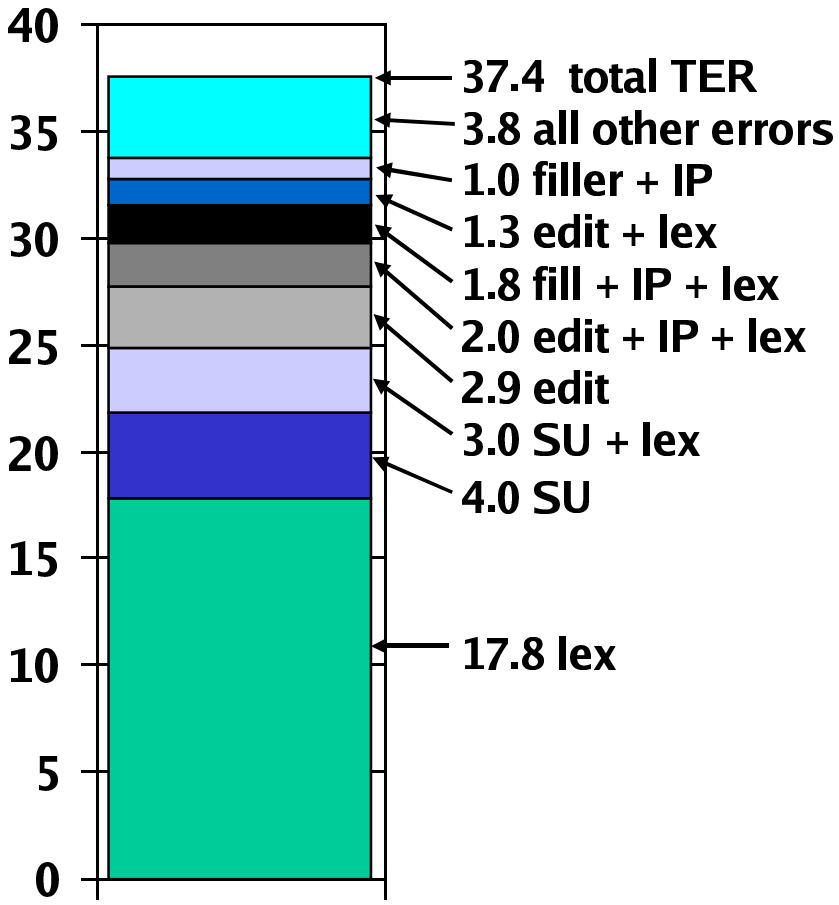
- Optimal SER for one metadata type in isolation does not lead to optimal TER
- Metadata SER is best understood in the context of an overall RT system TER
- Number of slots differs for each metadata type

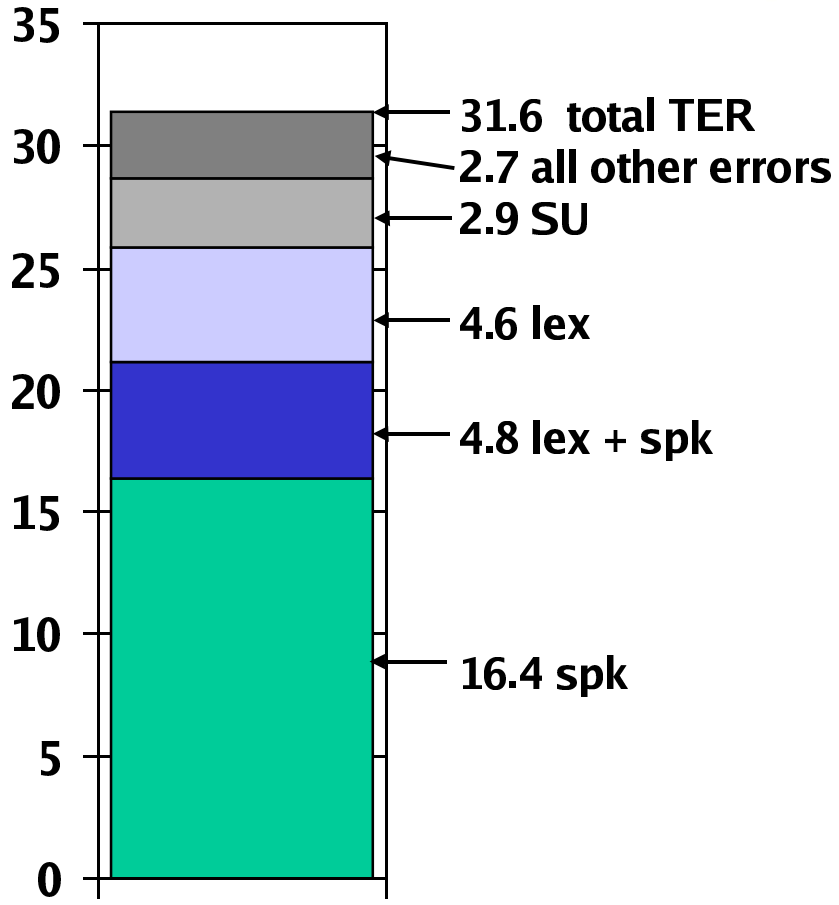
BBN TECHNOLOGIES
A Verizon Company

Motivations for RT TER Primacy

1. TER is the Edit Distance between the system output and the reference
   - Treats any type of token error as equal in application cost
2. TER encourages joint optimization of Metadata tasks
   A. Isolated development of metadata subsystems is suboptimal
3. TER promotes direct comparisons to the underlying STT WER
   A. Focuses attention upon the largest sources of error so that research effort can be directed most effectively

RT/MDE Error Distribution





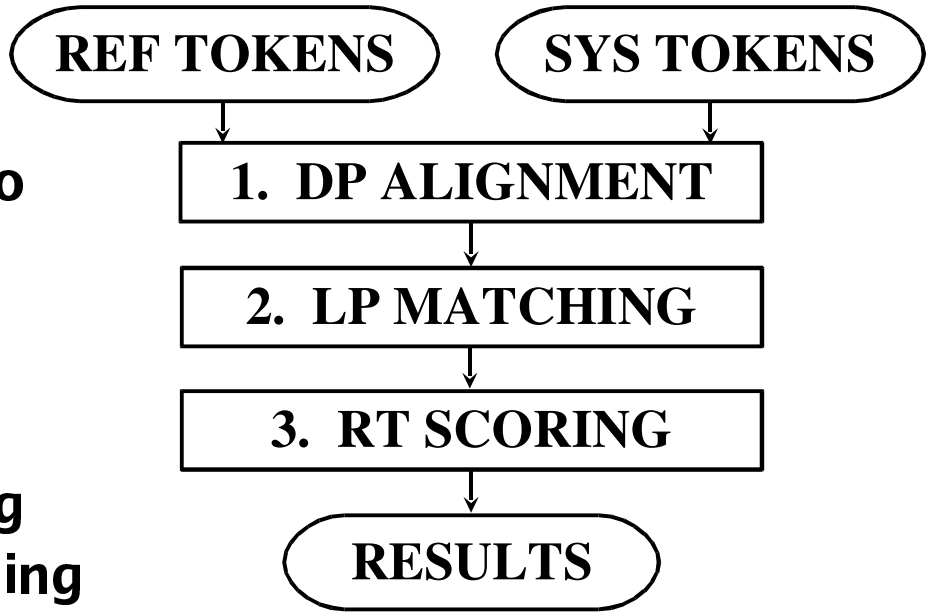
Lexeme and SU errors dominate CTS TER
Speaker and lexeme errors dominate BN TER

BBN TECHNOLOGIES
A Verizon Company

rteval - a Tool for RT Evaluation

***rteval* uses a 3-step procedure to calculate all RT and MDE scores from one common alignment**

1. **Align reference and system token sequences using Dynamic Programming (DP)**
2. **Match system speaker labels to reference speaker labels using Linear Programming (LP)**
3. **Score overall RT TER and each individual MDE SER using the same alignment and matching**



Note: Speaker boundaries are used in DP alignment since mapped speaker labels are not available until LP matching is complete

BBN TECHNOLOGIES
A Verizon Company

Benefits of *rteval*

- Employs well-known optimal algorithms for token alignment (DP) and speaker label matching (LP)
- Does not require any runtime tuning parameters or timing information
- Calculates scores for all conditions in one efficient pass
  - *rteval* requires < 2 min to score 1.5 hr BN or 3 hr CTS test sets on a Pentium 4 PC
- Written in Perl for portability, transparency, and easy modification
- Inputs and outputs are structured XML data
  - Scored results and alignments are browser-ready

Browsing *rteval* Results - a Demonstration

EARS The Blue Team
BBN LIMSI UPitt UW

Rich Transcription - Token Error Rate - Overall Scores												Expand
Rich Transcription - Token Error Rate - Episode Scores for	Episode: sw4386 - Channel: 1											Collapse
System	Nref	Nsys	Ncor	Nsub	Ndel	Nins	Nerr	%Cor	%Sub	%Del	%Ins	%TER
RT-03 Rich Transcription	366	339	227	101	38	11	150	62.02	27.60	10.38	3.01	40.98
RT1	366	339	253	72	38	11	121	69.13	19.67	10.38	3.01	33.06
Null Recognizer	366	0	0	0	366	0	366	0.00	0.00	100.00	0.00	100.00
Metadata - Slot Error Rate - Episode Scores												
System	Nref	Nsys	Ncor	Nsub	Ndel	Nins	Nerr	%Cor	%Sub	%Del	%Ins	%SER
Filler Detection	28	29	21	0	7	8	15	75.00	0.00	25.00	28.57	53.57
Edit Detection	23	0	0	0	23	0	23	0.00	0.00	100.00	0.00	100.00
IP Detection	35	24	18	0	17	6	23	51.43	0.00	48.57	17.14	65.71
Sentence Boundary Detection	35	38	25	0	10	13	23	71.43	0.00	28.57	37.14	65.71
Speaker Recognition	366	339	328	0	38	11	49	89.62	0.00	10.38	3.01	13.39

Ref Token	Sys Token	Ref Fill	Sys Fill	Ref Edit	Sys Edit	Ref IP	Sys IP	Ref Sent Bnd	Sys Sent Bnd	Ref Speaker	Hyp Speaker	Mapped Ref Speaker
increase.	freeze.							end	end	sw4386_1	1	sw4386_1
Do	Do									sw4386_1	1	sw4386_1
you	you									sw4386_1	1	sw4386_1
take	take									sw4386_1	1	sw4386_1
any	in									sw4386_1	1	sw4386_1
	the										1	sw4386_1
uh	um	filler	filler			beg	beg			sw4386_1	1	sw4386_1
um	um...	filler	filler			beg			end	sw4386_1	1	sw4386_1
i	I									sw4386_1	1	sw4386_1
won't	won't									sw4386_1	1	sw4386_1
say	take									sw4386_1	1	sw4386_1
steroids	steroids.								end	sw4386_1	1	sw4386_1
but	But									sw4386_1	1	sw4386_1
i'l	i			edit		end				sw4386_1	1	sw4386_1
i'll	i'll									sw4386_1	1	sw4386_1
say	say									sw4386_1	1	sw4386_1

Benefits of XML Data Format

- XML formats are extensible without breaking existing tools
- XML data can be automatically checked for correctness against its defining schema
- XML supports all languages included in Unicode 3.0
- XML parsers are built into in modern Web browsers

Online Resources for RT

- **Evaluation tool – *rteval v2.3***
  http://www.speech.bbn.com/ears/rteval_v2.3.zip
- **RT XML schema**
  http://www.speech.bbn.com/ears/rtxml_v2.3.xsd
- **RT Framework overview paper**
  http://www.speech.bbn.com/ears/Framework_for_Evaluating_Rich_Transcription_Technology.pdf

BBN TECHNOLOGIES
A Verizon Company